# Video Representation and Semantic Information Extraction using Ontology-based Approaches

**Ph.D. Synopsis Report**

*submitted by*

**Ashish Singh Patel**
(Roll No.:175000801)

*Under the supervision of*

**Dr. Vivek Tiwari**

***Department of Computer Science and Engineering,***
***International Institute of Information Technology, Naya Raipur***

**Dr. Muneendra Ojha and Prof. O. P. Vyas**

***Department of Information Technology,***
***Indian Institute of Information Technology, Allahabad***

**Department of Computer Science & Engineering**
**International Institute of Information Technology, Naya Raipur Raipur-493661**

**July, 2022**

# International Institute of Information Technology, Naya Raipur
# Declaration by the Student

**Title of the Thesis:** VIDEO REPRESENTATION AND SEMANTIC INFORMATION EXTRACTION USING ONTOLOGY-BASED APPROACHES

**I declare that,**

1. The work contained in the thesis is original and has been done by myself under the general supervision of my supervisors;
2. The work has not been submitted to any other institute for any degree or diploma;
3. I have followed the guidelines provided by the Institute in writing the thesis;
4. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute;
5. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis, giving their details in the references, and taking permission from the copyright owners of the sources, if necessary;
6. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

**Place:** Raipur

**Date:** July 8, 2022

Ashish Singh Patel
Roll No.: 175000801

डॉ.श्यामाप्रसाद मुखर्जी अंतरराष्ट्रीय सूचना प्रौद्योगिकी संस्थान
DR. SHYAMA PRASAD MUKHERJEE INTERNATIONAL
INSTITUTE OF INFORMATION TECHNOLOGY, NAYA RAIPUR
(A Joint Initiative of NTPC & Govt. of Chhattisgarh)

Email: iiitnr@iiitnr.ac.in
Tel: +91-771-2474040
Web: www.iiitnr.ac.in

# Declaration by the Supervisor(s)

This is to certify that the Thesis VIDEO REPRESENTATION AND SEMANTIC INFORMATION EXTRACTION USING ONTOLOGY-BASED APPROACHES, submitted by Mr. Ashish Singh Patel to the International Institute of Information Technology, Naya Raipur, is a record of bonafide research work under our supervision and we consider it worthy of consideration for the Synopsis of the Thesis to be submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy of the Institute.

(Supervisor)

Dr. Vivek Tiwari
Assistant Professor
Department of Computer Science and Engineering
International Institute of Information Technology, Naya Raipur

(Co-Supervisor)

Dr. Muneendra Ojha
Assistant Professor
Department of Information Technology
Indian Institute of Information Technology, Allahabad

(Co-Supervisor)

Prof. O. P. Vyas
Professor
Department of Information Technology
Indian Institute of Information Technology, Allahabad

Date: July 08, 2022

# Abstract

Due to the perceptual information about actions and events contained in video data, which is unstructured and hard to quantify, there is a semantic gap between computer interpretation and human understanding. Consequently, the majority of surveillance systems rely on a human agent to watch and identify suspicious situations. However, human agents are expensive and frequently afflicted by weariness and inattention, which may result in accidents. Additionally, storing and evaluating visual data is one of the most significant challenges. However, it can be resolved by encoding data in a machine-readable and machine-interpretable format, thereby facilitating semantic storage and retrieval. Moreover, the multifaceted benefits inspire researchers to investigate video representation in machine-readable and machine-interpretable formats for ease of storage, reasoning, and analysis in order to obtain semantic information. Various metadata leveraging Semantic Technologies for extracting and representing salient video information in a structured format are demonstrated in this thesis work which further helps to obtain semantic information of complicated events/activities/unusual events. This representation requires knowledge extraction, ontology engineering, representation in symbolic form, reasoning, and retrieval via querying.

The first section of the thesis discusses mid-level feature extraction, which helps for identifying object tracking information. It further utilized to identify crowd behavior of loitering and physical separation by employing a dynamic threshold proportional to the object size. In addition, a method is proposed to measure the real physical distance using the human height as a reference. The proposed technique performance has been bench-marked and validated with MOT datasets, PETS 2006 datasets, and UCF datasets, all of which indicate encouraging results. A full video representation framework is then proposed and demonstrated for identification of suspicious behavior in a parking space in the preceding section. A domain ontology is designed to represent the low-level features and the mid-level features respectively. The underlying inference rules are built using SWRL and queries were written to retrieve the information. In addition,a dataset has been created that consists numerous sequences of six labeled suspicious actions to evaluate proposed method. The framework is assessed by detecting unusual activity on the developed dataset, PETS 2006 and PETS 2014, and achieved state-of-the-art performance. In the third section, we describe the framework for storing, evaluating, and retrieving information within the setting of public place video surveillance. The framework is validated by detecting abandoned luggage on the AVSS 2007, PETS 2006, and ABODA datasets and obtaining a significant improvement of approximately 10 percent over existing approaches. The method for detecting odd behavior is substantially improved by incorporating context and other domain-specific data. The preceding contributing section includes the ontology-based multimodal analysis by merging unstructured video data and structured sensor data. Further demonstrates complicated event detection in illegal parking through the use of mobile surveillance vehicles. A use case has been taken for validation where a dataset contains illegal parking scenarios captured by a mobile surveillance car. An ontology is required for video representation; yet, it is a time-consuming task that demands domain expertise. The final section discusses automation of the data-driven and NLP-guided ontology construction process using target domain datasets. The presented ontology outperforms by more than 10% in representing visual data in each category on the Visual Genome dataset.

The presented techniques have many advantages over machine learning and computer vision based methods. In the end, it concludes that Semantic Web Technologies are proving an excellent toolchain for interpreting videos, leading to brand-new opportunities in the automatic representation of video scenes and identification of complicated events.

**Keywords:** *Video Surveillance, Semantic Technology, Video Representation, Ontology Engineering, SWRL, NLP*

## Contents


**1 Introduction** **1**

1.1 Motivation and objectives . . . . . 1

1.2 Research Challenges and Gap . . . . . 1

1.3 Research Contribution . . . . . 3

**2 Brief Literature Review** **3**

**3 Description of Research Work** **3**

3.1 Mid-level feature extraction for video representation and activity detection . . . . . 4

3.2 Video representation and event identification in parking lot using Semantic Web Technology . . . . . 6

3.3 VizPubs: A data-driven ontology and unusual activity detection in public Places . . . 8

3.4 Ontology-based data integration and multi-modal analysis . . . . . 11

3.5 Data-driven NLP guided ontology generation . . . . . 13

**4 Conclusion and Future Scope** **15**

**5 Organization of the Thesis** **15**

**6 Publications** **16**

# 1 Introduction

These days, video surveillance systems record a lot of footage, much of which goes unwatched because no one is monitoring the cameras. Over time, a human agent applies his or her knowledge, experience, and cognitive abilities to video data to comprehend it. Nevertheless, due to inherent physical limitations, humans experience fatigue and inattention, which can lead to an unnoticed but potentially hazardous situation. Therefore, an intelligent system capable of recognizing situations involving "unusual events of interest" is necessary [1]. The intelligent system must be able to access, process, store, and extract meaning from video content in order to bridge the semantic gap between machine output and human perception. Bridging the semantic gap necessitates a machine-readable and machine-interpretable representation of the extracted information involving spatial, temporal, and reasoning to produce high-level semantic information in the form of events, activities, various know-how, unusual events, and metadata.

## *1.1 Motivation and objectives*

According to recent data published by Cisco systems [2] on the growth rate of Video data, by 2021, 82% of internet traffic will consist of video and internet. Video surveillance data itself is projected to grow seven-fold. In addition to these developments, most surveillance data is generated but not publicly shared. Digital Cameras are capturing an unfathomable amount of Video data, the vast majority of which is unused, unanalyzed, and simply stored by organizations. This rapid increase in video content has necessitated the development of technologies to access, process, and store video content and extract meaningful information from it. However, there are many technical and domain-specific hurdles that must be overcome in order to accomplish the aforementioned goals. The objectives of this thesis work are listed as below:

- To develop techniques for extracting mid-level information from visual data.
- To propose techniques and frameworks for the machine-readable and machine-interpretable representation of unstructured visual data.
- To propose the method, tools, and framework for representing and storing semantic information present in visual data.
- To provide a method for identifying suspicious and unusual parking lot activity.
- To develop methods for identifying suspicious and out-of-the-ordinary behavior in public environments.
- Demonstrate information extraction, storage, and retrieval for video comprehension, activity detection, identification of unusual events, anomaly detection, and question answering while describing data expertise.
- Automate the process of developing ontologies to represent video data.

## *1.2 Research Challenges and Gap*

When it comes to reasoning and analysis, multimedia data is unstructured and must be converted into a form that computers can understand. By pre-processing the video data, extracting relevant information from it, representing it, and then retrieving it, the monitoring systems should be able to isolate specific pieces of information from the footage. The identification of higher-level features, also known as extremely complicated events, is accomplished by first determining lower-level features and then applying prior knowledge of the domain. Rajkumar et al. [3] enumerated a number of challenges for video analysis, including detecting events, mining concepts, analyzing motion, representing video, extracting features, and selecting the proper algorithm for video data-mining. The process of extracting, representing and retrieving high-level semantic information (such as events, actions, location of the event, and metadata) from the video in order to reach the human level of perception results in a semantic gap, which is illustrated in Fig. 1.

The research gap identified during this study can briefly be outlined as follows:

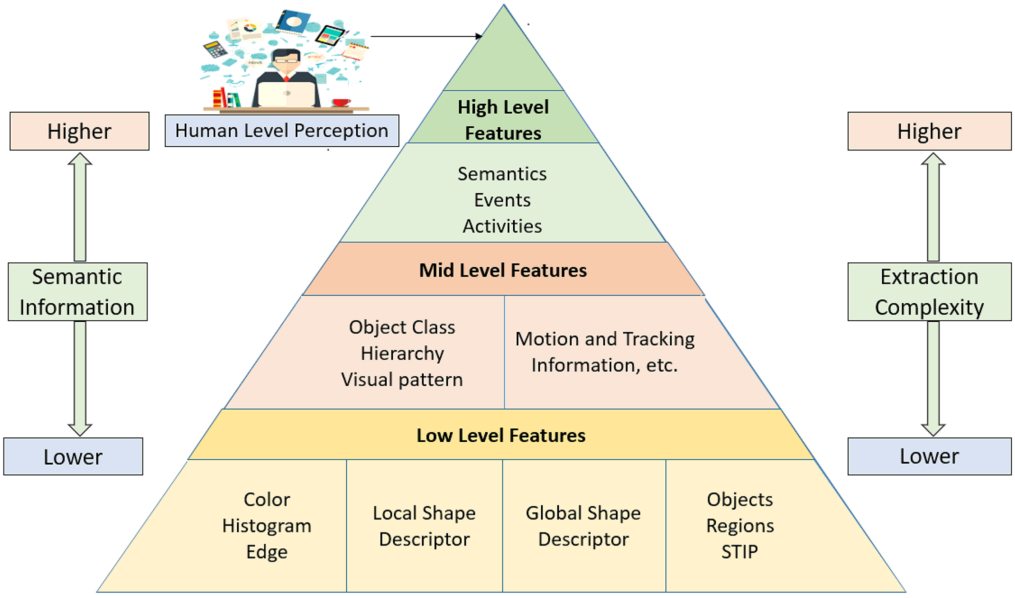


Figure 1: Low-Level information to high-level semantics

- The disparity between how a machine interprets something and how a person understands it, taking into account the human capacity for cognition, knowledge, and experience (semantic gap). The ability of humans to understand based on their cognition, knowledge, and experience is the primary cause of the semantic gap.
- The identification and extraction of high-level video semantics in the form of complicated events, activities, and information concerning the data.
- Low-level features do not offer any insight into the significance of the content that is being viewed.
- The majority of the existing methodologies do not utilize the multi-modal information or the contextual knowledge that may be available.
- Conventional methods in machine learning are data-driven; this requires a substantial quantity of training data and makes them challenging to implement when there is less of it, particularly in the event of abnormal occurrences.
- The majority of existing models are programmed to extract narrowly focussed data, which is frequently insufficient for a comprehensive analysis.

The semantic disparity can be closed by representing video in a format that is machine-readable and machine-interpretable, along with reasoning. This, in turn, calls for the extraction of low-level features and tracking information, as well as multi-modal data and the generation of an ontology. For this reason, it is necessary to develop methods that can extract, transform, and represent the unstructured information that can be found in the video format. The overall research problem addressed in this thesis is thus outlined as follows:

- Reducing the semantic gap between machine interpretation and human-level perception.
- Identification of semantics as complex events, activities, and data-related knowledge.
- Extraction of video content's semantics describing "Where? ", "What? ", "When? ", "Why? ", "Who?" and "How?" regarding the data.
- Retrieval of semantic information in the form of event, activity, diverse knowledge, and unexpected events.
- Storing the visual data's unstructured semantic information.

### *1.3 Research Contribution*

The present research on the thesis demonstrates how to extract and represent important information that may be present in video frames in a format that is machine-readable and machine-interpretable. This improves the ease of retrieval, processing, and storage, which is helped by multi-modal analysis and automatic ontology generation.The following is a synopsis of the significant contributions that this body of research has made:

- A Mid-Level Features extraction approach for video representation and activity detection.
- An algorithm that improves the localization of an object in a video by employing a data association approach to estimate object tracks and alleviate Identity Switches.
- A tracking algorithm identifies suspicious activity ("loitering") in a video and assigns a severity level (high/medium/low).
- Frame level representation of Video scene in RDF, which saves storage space, facilitates, reasoning and efficient information retrieval.
- Deriving relationships between the objects in an image using SWRL. A novel approach for object tracking is also proposed based on SWRL and Description Logics.
- A framework to detect events and represent surveillance data in a machine-readable and machine-interpretable format for reasoning, storing, querying, and retrieving. An activity detection method using SPARQL and RDF graph.
- The extracted information/structure representation in VizPubs ontology and queried using DL and SPARQL. Identify the activities and semantic information with the use case of abandoned luggage detection.
- An ontology for representing and storing roadside surveillance data as knowledge graphs, identifying illegally parked vehicles, and public place representation.
- A class and object extraction technique based on domain data and NLP-guided approach. Semantic refinement and spatial relation extraction using frequency-based relation ranking inference, and reasoning. A novel ontology (VizITS) for the Intelligent Transportation System (ITS) domain, as a running example for the proposed approach.

## 2 Brief Literature Review

Methods for extraction of features have been widely investigated in past and can be found in literature. These methods cover the extraction of low-level features as well as mid-level information, both of which are necessary for the generation of high-level video semantic information. The extraction is accomplished through a variety of strategies that make use of computer vision and machine learning. Since the development of Deep Learning [19] has given rise to a new facet in the generation of high-level semantic information, it has also been investigated in depth from the point of view of activity detection and abnormal activity detection. Direct methods that make use of high-level semantic information tend to have lower accuracy and can only be applied to specific contexts, whereas indirect methods that make use of low-level and mid-level features can be applied to a wider variety of contexts. In addition, direct approaches have to be meticulously prepared for specific example scenarios. Methods that concentrate on the most important aspects have the ability to generalize their results and enjoy wider acceptance. Table 1 provides a concise summary of the research done using a variety of methodologies and presents the findings.

## 3 Description of Research Work

This thesis work presents emerging Semantic technologies for extracting, representing, and reasoning over semantic knowledge. In addition to this, ontology-based multi-modal analysis is illustrated through the use of automatic ontology generation.

Table 1: A brief taxonomy of the approaches related to the proposed thesis work

| Work | Related Methods | Description | Comments |
|---|---|---|---|
| Trajectory based Activity Analysis | Trajectory analysis [4], DiMoLIF [5] | Event Analysis by classifying object trajectory | Can be used for tracking objects for the purpose of event detection and tracking based activity analysis |
| Real-time Object Tracking | associating tracks with detections, [6], [7] | Generate tracks of the objects in the video | Used for event analysis, both normal and unusual |
| Activity Detection | P3D[8], [9], TRN [10], TSN [11] | Deep Learning methods which utilize the DNN architecture, Trimmed Activity recognition | Very computational complex, Activity localization is a major challenge as well as unusual behavior recognition |
| Abnormal and Complex Events Detection | Two-Stream DSTAE [1], AEP [12] | Deep Learning based and Hybrid Approaches that employ Deep Learning | Suffers from lack of training examples for abnormal events, more complex events are difficult to detect due to large intra variation. |
| Video Ontology | LSCOM [13], VidOnt[14], MPEG-7 [15] | Ontology Engineering approaches for multimedia | Need to develop/automate the ontology generation |
| Activity Recognition using semantic technologies | OVIS [16], [17],[18] | Activity/Event Identification using SPARQL, SWRL, and inference | Semantic Event analysis, Inference and Reasoning |

### *3.1 Mid-level feature extraction for video representation and activity detection*[1]

In order to maintain public safety and cooperation with law enforcement, an intelligent system is required to monitor public spaces in search of odd behavior. However, detecting such actions requires information both spatial and temporal, which includes the interaction of many objects inside the same frame as well as subsequent temporal frames that require reliable tracking of the multiple items.

The work that is being proposed to extract mid-level features has been divided up into four different categories, which include object detection and tracking, loitering classification, adaptive threshold-based, and actual physical distance estimation. The following is a condensed version of the contribution that this work makes:

- A method for improving the localization of an object in a video by using a data association approach to estimate the tracks of the object in a way that reduces the occurrence of ID Sw.
- A tracking algorithm that can recognize suspicious behavior (i.e. "loitering" in the present study) in a video and assign a severity rating (high, medium, or low) to it using an online method.
- A method for determining the actual physical distance in a crowd by employing reference thresholding.
- A method for estimating the actual physical distance using reference objects.

The framework of the proposed tracking-based activity detection and behavior analysis is subdivided into three parts; object detection, object tracking, and crowd behavior analysis as depicted in Fig. 2.

[1]Patel, Ashish Singh, Ranjana Vyas, O. P. Vyas, Muneendra Ojha, and Vivek Tiwari, *Motion-compensated online object tracking for activity detection and crowd behavior analysis*,The Visual Computer, Springer, 1-21, IF:2.6, (2022).

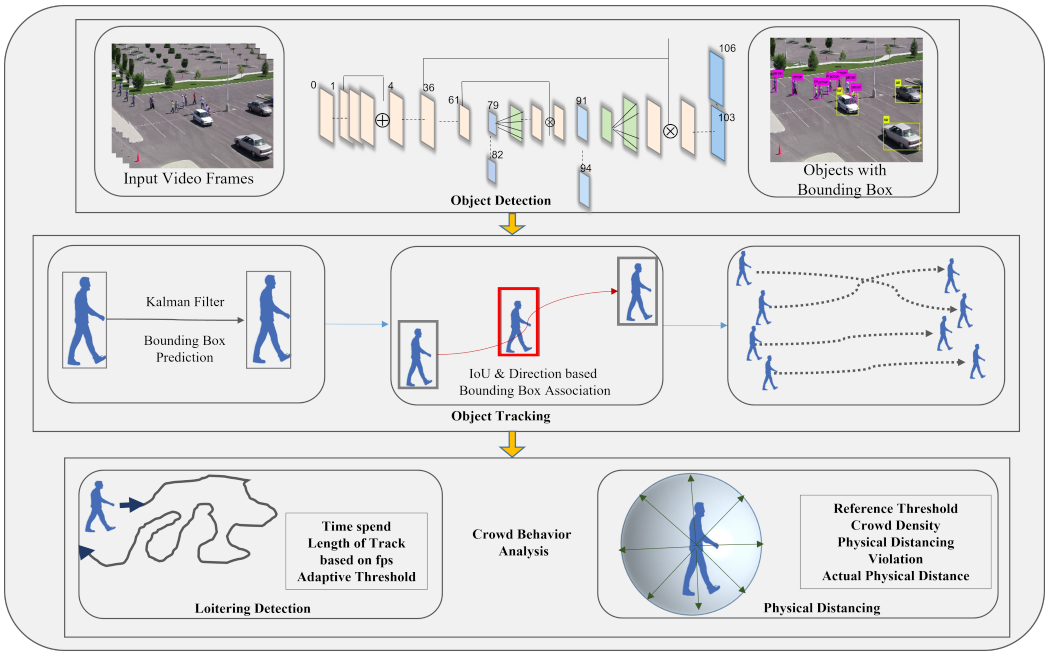


Figure 2: A proposed framework for activity detection and crowd behavior analysis

Fig. 2 is divided in three parts. A layered model is included in the first part of the image, which presents the details of the object detection method known as YOLOv3 [20]. The information on the tracking algorithm is depicted in second part, which consists of prediction followed by data association. Tracking information is also used for the analysis of crowd behavior, specifically to determine activities such as loitering and physical distancing, with the actual physical distance measurement being presented in the third part of the image. Key elements and parameters used in crowd behavior analysis are also provided under each stage. The essential feature includes the adaptive thresholding techniques for assigning a severity level and adaptive threshold for physical distancing based on the size of the bounding box with the real physical distance by considering the height of the human as a reference in the case of adults, as shown in Fig. 3 using Eq. 1, 2, and 3.

$$PD_{ij} = \sum_{y=o_{iy}}^{o_{jy}} \frac{\sqrt{\Delta x^2 + \Delta y^2} * 1.8}{(y - y_t) * 2} \tag{1}$$

$$x = \frac{y - c}{m} \tag{2}$$

$$m = \frac{o_{iy} - o_{jy}}{o_{ix} - o_{jx}} \tag{3}$$

The detection of the behavior of loitering, which makes use of the tracking information as a baseline, is one way that the efficacy of the object tracking approach can be proved. As a result, the suggested method is analyzed using three fundamental stages when viewed from the standpoint of activity detection. These stages include object identification, object tracking, and loitering detection. The PETS dataset is used here. This is a benchmark dataset for evaluating the performance of object tracking, and it includes a variety of actions including loitering. In addition, the effectiveness of the proposed tracking and physical distancing method, which includes the estimation of actual physical distance, has been demonstrated by validation on MOT benchmark datasets [21] (PETS09-S2L2, ETH Crossing, AVG-TownCentre, TUD-Stadmitte, and Venice-2), which were used to test the method. Since the MOT benchmark datasets already contain object detections, any potential bias in the method of object detection used to measure tracking performance has been eliminated. A comparison is also

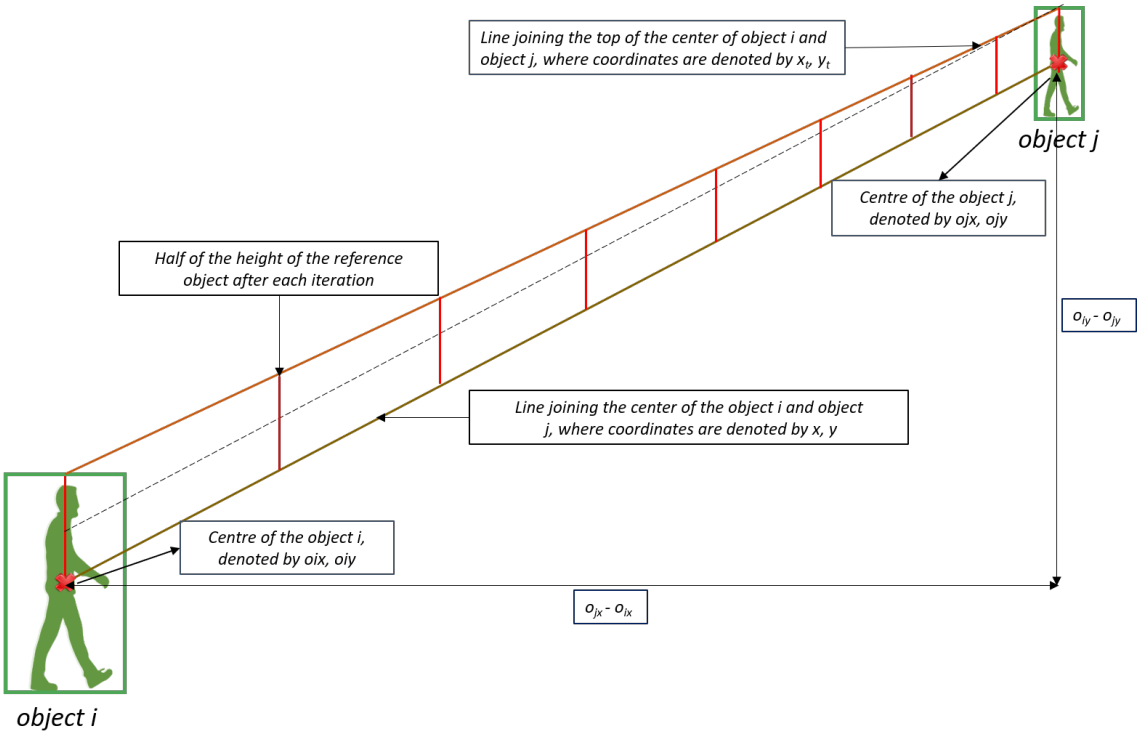
Line joining the top of the center of object i and object j, where coordinates are denoted by $x_t$, $y_t$
object j
Centre of the object j, denoted by ojx, ojy
Half of the height of the reference object after each iteration
$o_{iy}$ - $o_{jy}$
Line joining the center of the object i and object j, where coordinates are denoted by x, y
Centre of the object i, denoted by oix, oiy
$o_{jx}$ - $o_{ix}$
object i

Figure 3: Estimation of actual physical distance by using reference object

performed with the recent object tracking approaches on the dataset D3 (PNNL2 [7]) in Table 2. The proposed strategy not only improves accuracy but also brings the overall number of ID Sw down by a significant amount.

Table 2: Performance of the proposed algorithm on dataset (PNNL2 [7])

| **Approach** | **Evaluation Metric** | | | |
|---|---|---|---|---|
| | **MOTA** | **MOTP** | **MT** | **ID Sw** |
| CMOT [22] | 80.7 | 58 | 85.9 | 6 |
| IHTLS [23] | 78.8 | 57 | 87.5 | 5 |
| GMMCP [7] | 87.6 | 58 | 90.6 | 7 |
| SORT [6] | 84.6 | 59 | 88.3 | 6 |
| Proposed Work | 91.4 | 61 | 92.5 | 3 |

The approach can handle short-term object occlusion, detection errors, and ID Sw. The generated accurate tracks of the objects is used for analyzing crowd behavior in public places. The proposed work opens up a plethora of use-cases for detecting usual and unusual activities/events by using the tracking information of the objects in a surveillance environment. The use of context and tracking information can be vital in detecting activities in real-time and complex scenarios.

### *3.2 Video representation and event identification in parking lot using Semantic Web Technology*[2]

The semantic information present in video needs to be extracted, transformed, and linked with domain knowledge to acquire interpretation capabilities through software agents [24]. This requires strong reasoning and analytical support to be able to detect an event, especially of anomalous nature and bridge

[2]Patel, Ashish Singh, Giovanni Merlino, Dario Bruneo, Antonio Puliafito, O. P. Vyas, and Muneendra Ojha, *"Video representation and suspicious event detection using semantic technologies"*, Semantic Web, IOS Press, 12(3), 467-491, IF:3.5, (2021)

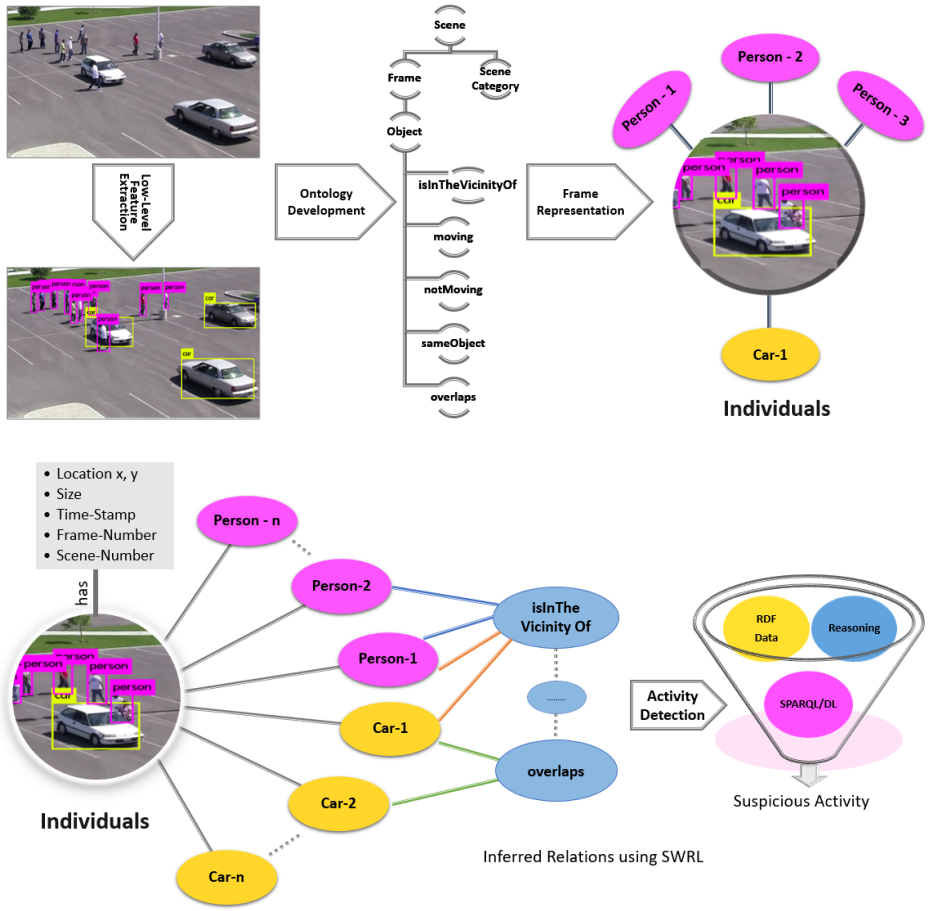


Figure 4: The workflow of the proposed approach

the semantic gap between machine interpretation and human perception. Semantic Web Technology, is, therefore leveraged to fill this gap by creating domain ontology, which is effective in representing high-level semantics present in the video. The list of the tasks accomplished in this work are as follows:

- Frame level representation of Video scene in RDF, which saves storage space, facilitates reasoning and efficient information retrieval.
- Deriving relationships between the objects in an image using SWRL i. e. reasoning over video.
- Activity detection using SPARQL: Once all the information is represented in RDF graphs, activity / events can be recognised and retrieved by formulating SPARQL.
- A video dataset which consists of six different trimmed localized activities in smart parking scenario totaling 92 videos.
- An ontology is developed capable of representing parking lot.
- A novel approach for object tracking is also proposed based on SWRL and Description Logics.

In this work, a method to detect suspicious events occurring inside the parking lot to monitor the safety of parked vehicles is proposed. The object along with its bounding-box is extracted using Yolov3 [20]. Mid-level features consist of class hierarchy of the objects and relationships between the objects. A detailed workflow is shown in Figure 4 wherein extraction of low-level features, ontology development, and frame representation, with the generation of relationship and detection of an activity. The constructed ontology represents the spatio-temporal relations between objects in a video scene. The features of a frame are represented using ontology as data properties. In contrast to data properties, the object properties are referred to as mid-level features.

As shown in Fig. 4, each object of the frame is represented as an individual. In Fig. 4, the mid-level features, i.e., the relations between the objects present in same frame and temporal frames are

Table 3: Description of Activity Dataset

| Activity Description | Video Count | Suspicious | Alertness Level |
|---|---|---|---|
| Person walking around the car | 10 | Yes | High |
| Person touching the car | 15 | Potentially | Low |
| Person looking inside the car | 24 | Yes | High |
| Person attempting to damage the car | 17 | Yes | High |
| Group of people passing nearby car | 11 | Yes | Low |
| Trying to open the door of the car | 15 | Potentially | High |

Table 4: Performance of inferring the relations in a video scene

| Relation | Total Relations | Precision (%) | Recall (%) |
|---|---|---|---|
| overlaps | 2436 | 97.63 | 98.59 |
| isInTheVicinityOf | 380 | 98.12 | 99.43 |
| sameObject | 60475 | 81.72 | 97.45 |
| moving | 5429 | 98.36 | 99.37 |
| notMoving | 1896 | 96.03 | 98.71 |

inferred using SWRL rules between the individuals. The unusual activity detection is performed by formulating Description Logics (DL) and SPARQL queries as a information retrieval task. An activity dataset[3] that contains six unusual activities, with a total of 92 videos is created. Details of the activity dataset is shown in Table 3.

The proposed work is evaluated by measuring the accuracy of the inferred relations and activities. Accuracy of SPARQL query directly depends on inferred relationship as it extracts presence of relationship in the temporal domain. The accuracy of inferred relations is measured on open parking dataset [7] and the accuracy of activities on the developed dataset containing six activities in parking lot environment. The accuracy of the inferred relations is measured by calculating precision and recall of relations by forming a confusion matrix shown in Table 4. The matrix is calculated by considering relations among 7325 objects of open parking dataset [7]. Each frame has 14 objects on an average, and the number of frames per second (fps) of the video is 15. In total 70616 relations (*overlaps, isInTheVicinityOf, sameObject, moving and notMoving*) exists between the objects, of which the inferred *sameObject* relation has more number of false positives because of close objects and object occlusion. Thus, proposed framework performs quite well in inferring the complex spatio-temporal relations and interactions between the objects. Activity recognition is performed by executing SPARQL queries tailored explicitly for a particular activity. The number of true alarm, false alarm and missed alarm are recorded to calculate the accuracy of the approach. The results are shown in Table 4.

The proposed work demonstrated the potential in leveraging semantic web technology for activity detection, especially useful in scenarios featuring lack of training data and limited computing resources. By detecting these sub-activities, more complex (and atypical) activities can be derived.

### *3.3 VizPubs: A data-driven ontology and unusual activity detection in public Places*[4]

A framework to detect events and represent surveillance data in a machine-readable and machine-interpretable format for reasoning, storing, querying, and retrieving is presented. Additionally, a

[3] https://github.com/aspdr/Video-Representation-and-Suspicious-Event-Detection-using-Semantic-Technologies/tree/master/Trimmed%20Activity%20Dataset

[4] Ashish Singh Patel, and Muneendra Ojha, O P Vyas, *"Accident Detection Using Time-Distributed Model in Videos"*, In Proceedings of Fifth International Congress on Information and Communication Technology, pp. 214-223. Springer, Singapore, 2021.

Table 5: Performance of the various activities by the framework

| **Activity Description** | **Total** | **True Alarm** | False Alarm | **Misses** | **True Alarm (%)** |
|---|---|---|---|---|---|
| Person walking around the car | 10 | 8 | 0 | 2 | 80 |
| Person touching the car | 17 | 13 | 2 | 2 | 76.67 |
| Person looking inside the car | 26 | 20 | 2 | 4 | 76.92 |
| Person attempting to damage the car | 17 | 15 | 0 | 2 | 88.23 |
| Group of people passing nearby car | 11 | 6 | 0 | 5 | 54.55 |
| Trying to open the door of the car | 15 | 12 | 0 | 3 | 80 |

method to detect complex events occurring in public places is proposed. The overall workflow of the approach is presented in Fig. 5. The task accomplished in this work are summarized as follows:

- A framework to detect events and represent surveillance data in a machine-readable and machine-interpretable format for reasoning, storing, querying, and retrieving.
- Ontology is developed to capture the concepts of public places.
- Extracted information is represented in the developed ontology VizPubs.
- Reasoning is performed using one of the reasoner (Pellet) by formulating SWRL rule.
- Structured representation in knowledge graph is queried using DL and SPARQL queries to identify the activities and semantic information with the use-case of abandoned luggage detection.

Firstly, the objects and tracking information is extracted from the surveillance data. Secondly, the ontology is developed to capture the concepts of public places. The extracted information is represented in the developed ontology *VizPubs*. Furthermore, reasoning is performed using one of the reasoner (Pellet) by formulating SWRL rules. In the final step the structured representation in knowledge graph is queried using DL and SPARQL queries to identify the activities and semantic information with the use-case of abandoned luggage detection. The *notMoving* property is established between

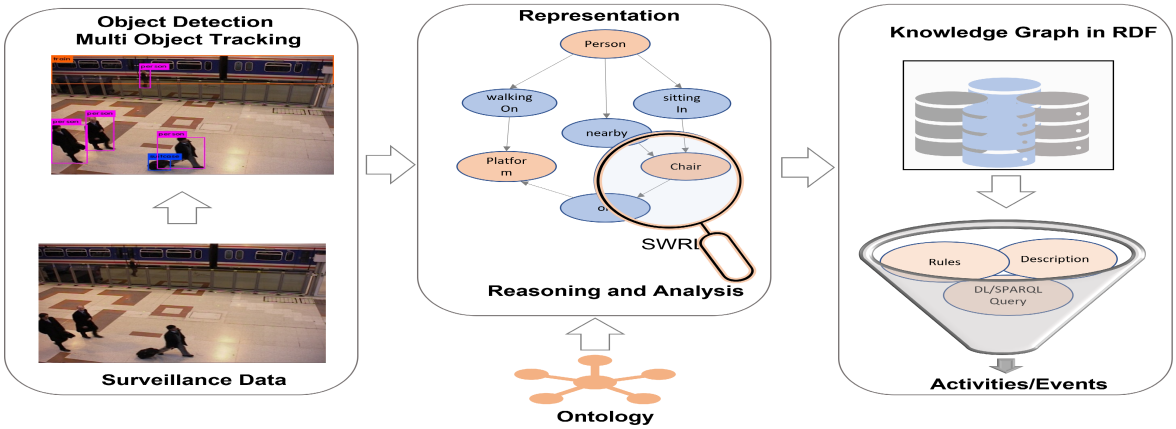


Figure 5: Workflow of the proposed unusual activity detection approach

two individual by comparing the bounding box location of the center of the bounding box with all the objects in the previous frame.

***notMoving***: The *notMoving* property is established between two individual by comparing the bounding box location of the center of the bounding box with all the objects in the previous frame.

Rule 1

```
Object(?i1)∧Object(?i2)∧frameId(?i1,?f1)∧frameId(?i2,?f2)∧
swrlb:add(?a1,?a,1)∧swrlb:equal(?a1,?f2)∧objectX(?i1,?x1)∧
objectY(?i2,?x2)∧swrlb:subtract(?xs,?x1,xsT)∧swrlb:add(?xa,
?x1,xaT)∧swrlb:greaterThan(?x2,?xs∧swrlb:lessThan(?x2,?xa)∧
objectType(?i1,?t)∧objectType(?i2,?t2)∧swrlb:equal(?t,?t2)∧
objectY(?i1,?y1)∧objectY(?i2,?y2)∧swrlb:add(?ya,?y1,yaT)∧
swrlb:subtract(?ys,?y1,ysT)∧swrlb:greaterThan(?y2,?ys)∧
swrlb:lessThan(?y2,?ya)∧swrlb:Equal(?y1,?y2)∧swrlb:Equal(
?x1,?x2)⇒notMoving(?i1,?i)
```

The abandoned luggage detection is formulated using Description Logic (DL) and formalized as: $\exists$ (Luggage A) $\sqcap$ (notMoving hasValue:true) $\sqcap$ (trackLength hasValue$\geq$ 150) $\sqcap$ $\nexists$ (Luggage A) (near $\exists$ (Person P)

The efficacy of the *VizPubs* is demonstrated by representing the data of public surveillance in three public place environment; road, park, and street environment. The presented ontology contains 126 concepts, including 54 classes, 14 attributes mapped as data properties, and 34 object properties (i.e., relations between classes). The class is extracted using a domain dataset, which is to be represented as knowledge graph (SPO triplet). We identified three distinct scenarios from the Visual Genome dataset of public places and evaluated the coverage of relations in the presented ontology. The three basic scenarios are "road environment", "street", and "park". The proposed ontology-based framework is evaluated by detecting abondoned luggage detection in PETS 2006 dataset and ABODA[5], and AVSS 2007 dataset. Table 7 demonstrate the performance of the proposed approach on AVSS 2007 and PETS 2006 dataset and results are compared with the DL based approach [25]. The results demonstrates the proposed approach performs better as compared CNN-based approach[25] as the duration is also considered in the proposed approach.

Table 6: Evaluation of proposed ontology on concepts of Visual Genome dataset [26]

| **Scene** | **Region (%)** | **Attribute (%)** | **Relationship (%)** | **Total (%)** |
|---|---|---|---|---|
| Road | 58.75 | 54.2 | 88.18 | 65.41 |
| Street | 47.22 | 76.24 | 91.74 | 74.28 |
| Park | 42.34 | 67.65 | 87.64 | 69 |

Table 7: Precision and recall of Abandoned Luggage Detection

| **Dataset** | **Method** | **Precision** | **Recall** |
|---|---|---|---|
| AVSS 2007[2] | Deep Learning [25] | 97.5 | 66.1 |
| | Proposed Work | 99.1 | 78.3 |
| PETS 2006 [27] | Deep Learning [25] | 95.6 | 83.6 |
| | Proposed Work | 98.1 | 91.3 |

The visual information present in each frame/image is extracted and represented using *VizPubs* ontology, which provides seamless interoperability and integration with strong reasoning capability. The proposed ontology-based framework is developed to capture and analyze the entire public place environment using surveillance systems. The proposed novel framework can represent entire salient information that can be queried to retrieve semantic information and metadata in the form of complex events, activities, and abnormal behavior.

[5] https://github.com/kevinlin311tw/ABODA

### 3.4 Ontology-based data integration and multi-modal analysis[6]

The surveillance systems have various IoT devices that generate the data in multiple formats, which predominantly contain textual data or multimedia data. The textual information includes the string data or the numeric data from the sensors such as temperature, GPS, etc., which can be directly quantified and represented. The multimedia data from the camera and audio recording systems contains the images/videos and audio data, which are unstructured and cannot be represented in machine-readable and machine-interpretable. In this work, a dynamic ontology-based framework is proposed to identify illegal parked vehicles using mobile surveillance leveraging 5G technology. The tasks accomplished in this work are summarized as below:

- An architecture leveraging 5G technology for dynamic monitoring and management of traffic in real-time for ITS environment.
- An ontology for representing and storing roadside surveillance data as knowledge graphs.
- A data integration approach that integrates sensory data with video data to perform complex reasoning.
- A framework for information storage and representation about the vehicles present across the city by presenting the concept of mobile surveillance. The information retrieval is performed using DL and SPARQL queries.
- A novel ontology-based algorithm to identify illegally parked vehicles.
- A real-time surveillance dataset containing illegally parked vehicles.

It is proposed to solve this crucial problem by using a mobile vehicle as an agent, Smart Bike Sharing System (SBSS) in this case, which captures the Video, location, and speed through various sensors. The surveillance data is captured through a mobile vehicle which continuously captures the video feed and sensor data (location, speed, and timestamp) which is further connected to the edge server with 5G network. At the edge server, the visual data is processed to detect the objects, generate the tracking information, Automatic Number Plate Recognition (ANPR) to identify the registration number of the vehicle and detect the lane on the road for identifying the illegally parked vehicles. The overall workflow of the proposed ontology-based approach is shown in Fig. 6. The data from the various sources (GPS, speedometer, and camera) is gathered and integrated using the developed ontology. The reasoning efficacy of the framework is validated by identifying illegally parked vehicles that are parked on the roadside by formulating SPARQL and DL Query. The SPARQL queries are formulated to retrieve semantic information in terms of classification, noteworthy information, metadata, complex events, question answering. The SPARQL query shown below outputs the instances (illegal parking) where individual *i1* is instance of class *Vehicle* and individual *l1* instance of class *Lane* such that *i1* is *below*

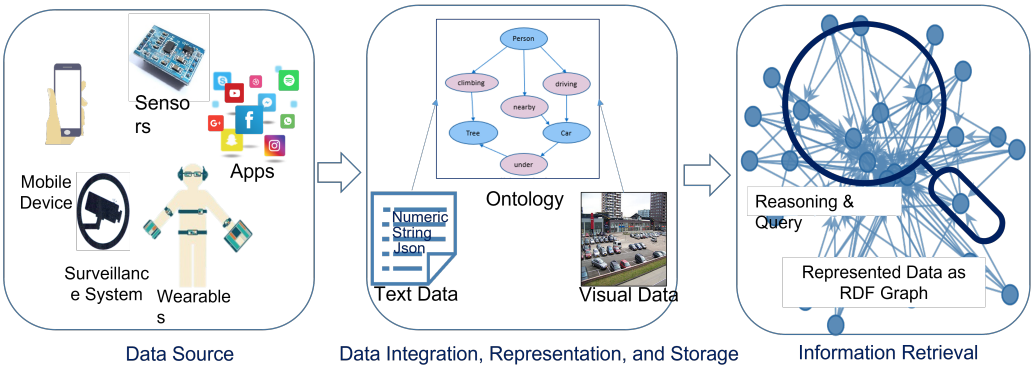


Figure 6: Workflow of the proposed ontology-based data integration, representation and retrieval approach

[6]Patel, Ashish Singh, Muneendra Ojha, O. P. Vyas, and Ranjana Vyas, *"Ontology-based multi-agent smart bike sharing system (SBSS)"*, In 2018 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 417-422. IEEE, 2018.

*l1* and *i1* has value *hasvehicleStationary* set to *true*. Initial lines of the query define the predefined prefixes, where *ipark* is prefix corresponding to our proposed ontology.

SPARQL Query

```
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
PREFIX time:<http://www.w3.org/2006/time#>
PREFIX date:<https://www.w3.org/TR/owl-time/#>
PREFIX ipark:<http://www.semanticweb.org/patel
 /ontologies/2021/4/ITS>
SELECT   distinct   ?i1   where{
 ?i1 rdfs: subclassOf Vehicle,
 ?l1 rdfs: subclassOf Lane.
 ?i1  ipark: below ?l1.
 ?i1  ipark:hasvehicleStationary   "true" .
}
```

A dataset of illegally parked vehicles in the city in real-time is recorded. The dataset consists of 7 different video sequences recorded in Naya Raipur Location with trip data at varying speeds and multiple scenarios/locations as mentioned in Table 8. The setup consists of two mobile cameras

Table 8: Description of the dataset containing illegally parked vehicles

| **Sequence** | **Duration (secs)** | **Scenario** |
|---|---|---|
| SEQ-NR-01-01 | 48 | capture stationary vehicle on the roadside moving at average speed of 8kmph at location 1 |
| SEQ-NR-01-02 | 23 | capture front view of the stationary vehicle on the roadside, moving at average speed of 20kmph at location 1 |
| SEQ-NR-01-03 | 21 | capture front view of the stationary vehicle on the roadside, moving at average speed of 18kmph at location 1 |
| SEQ-NR-01-04 | 28 | capture stationary vehicle on the roadside moving at average speed of 10 kmph at location 1 |
| SEQ-NR-02-01 | 24 | capture stationary vehicle on the roadside moving at average speed of 10 kmph at location 2 |
| SEQ-NR-03-01 | 24 | capture the side view of the stationary vehicle on the roadside moving at average speed of 24kmph at location 3 |
| SEQ-NR-03-02 | 18 | captures the front view of the stationary vehicle on the roadside moving at average speed of 24kmph at location 3 |

(smartphone) installed on a vehicle where the data is integrated using the timestamp of the Video. The dataset consists of 7 different video sequences recorded in Naya Raipur Location with trip data at varying speeds and multiple scenarios/locations with two cameras viewpoint. The illegal parking event identifies if the vehicle is stationary using the SPARQL query and SWRL rules. However, some false positive cases occur due to incorrect estimation of the length of the vehicle when it is far from the camera.

The surveillance data is captured using mobile surveillance vehicle which continuously sends the data to the edge for further processing. The track-level information is extracted and represented using

Table 9: Evaluation of Object Detection, Illegal Parking and ANPR

| Parameter | Evaluation Metric | | |
|---|---|---|---|
| | **Precision (%)** | **Recall (%)** | **Comments** |
| Object Detection | 92.3 | 94.9 | Vehicle is not detected when it is covered less than half of it |
| Illegal Parking | 88.4 | 89.3 | Vehicle length is incorrectly estimated when vehicle is far, high FP |
| ANPR | 89.4 | 88.7 | Depends on the vehicle speed and capture quality, and speed |

*ipark* ontology, which provides seamless interoperability and integration among heterogeneous sensor devices. The proposed ontology-based framework is developed to capture and analyze the entire road environment using mobile vehicles with installed sensors to detect the illegal parked vehicles.

### *3.5 Data-driven NLP guided ontology generation*[7]

Ontologies require domain-specific expertise to capture the domain concepts finitely and efficiently. Furthermore, finding a domain expert is a crucial task and requires expensive resources. Ontology-based approaches for event detection are also proposed by Tani et al. [17] and Sobhani et al. [18], which require the construction of ontology for representing the video. The list of the tasks accomplished in this work is summarized as follows:

- A novel *class extraction* technique based on domain data by identifying the objects present in the video.
- Object properties extraction using NLP-guided approach using textbase of the domain.
- Semantic refinement using an image captioning dataset, based on frequency-based relation ranking to find the most likely relations.
- Rules for *spatial relation extraction using inference* and reasoning.
- A novel data-driven *video surveillance ontology (VizITS) for the Intelligent Transportation System (ITS)* domain, as a running example for the proposed approach.
- Information retrieval mechanisms in the form of events, metadata, and other semantic properties using DL and SPARQL queries.

In this work, an approach for ontology development, video representation and information retrieval is proposed. This ontology development process is two-fold; a domain-dependent portion, driven by data, and a domain-independent one defining the properties of the data. The data-driven portion is constructed by obtaining class, object properties, and data properties of the ontology using target domain data and textual domain data for domain-dependent parts. Extraction step involves the extraction of concepts and entities for ontology. The class extraction is done using the objects detected in each frame of video, and then NLP is used for relationship extraction. Furthermore, the relations which are shared between a given subject and object are identified by calculating the intersection between the predicates linked to the subject and the object ($R_S \cap R_O$). $R_S$ refers to a set of predicates linked to the subject, while $R_O$ refers to a set of predicates linked to an object. In common relations, the most occurring one is picked by calculating its probability using eq. 4, denoted by P. The probability is defined as the ratio between the frequency of a relation, i.e., intersection as above indicated, and the sum of frequencies for all relations. The frequency of occurrence of a relation is denoted by $f_i$, where n is the total number of relations.

[7]Patel, Ashish Singh, Giovanni Merlino, Antonio Puliafito, Ranjana Vyas, O. P. Vyas, Muneendra Ojha and Vivek Tiwari, *"An NLP-guided ontology development and refinement approach to represent and query visual information"*, Expert Systems with applications, Elsevier, IF: 8.6 (Minor Revision)

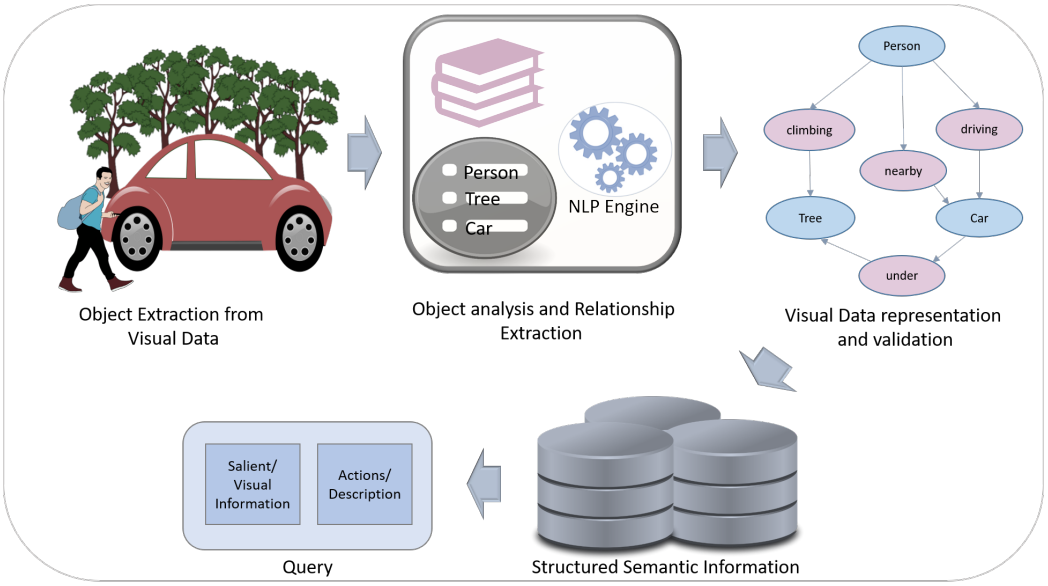


Figure 7: The proposed ontology development, data representation and information retrieval approach

Table 10: Comparison of the concepts in the ontology with existing video ontology

| **Ontology** | **High-level Semantics (%)** | **Mid-level** (%) | **Low-Level (%)** | **Total (%)** |
|---|---|---|---|---|
| LSCOM [13] | 18.6 | None | None | 18.6 |
| Ontology for Media Resources [9] | 16.6 | 61.4 | 12.8 | 30.2 |
| VidOnt [14] | 61.7 | 67.3 | 43.14 | 57.8 |
| VizITS | 57.2 | 71.4 | 68.2 | 65.7 |

$$P = \frac{f_i}{\sum_{i=1}^{n} f_i} \tag{4}$$

If a particular relationship does not occur with any other class, then we calculate the most probable relation inside the union of relations ($R_S \cup R_O$) using the same probability formula 4 but with the union as a set. An overview of model architecture used for ontology generation, representation and retrieval is shown in Fig. 7. The data represented in machine-readable machine-interpretable form is queried to extract visual and semantic information using SPARQL and DL query.

The presented ontology VizITS[8] contains 65 classes, 40 attributes mapped as data properties, and 63 object properties (i.e., relations between classes). The concepts of the proposed ontology are compared with the existing multimedia ontology LSCOM [13], EBU Core, Ontology for Media Resources and VidOnt [14]. The ontologies contain concepts for describing low-level features, spatio-temporal relations, administrative data (contains the information like sampling rate, frames per second, etc.), and high-level semantic information. Furthermore, these ontologies are built for the purpose of annotation and reasoning, but lack capabilities to represent visual and noteworthy frame-level information including high-level scene semantics extensively. Table 10 demonstrates the effectiveness of our presented ontology in representing the information present in the video belonging to a certain domain. The proposed framework can also be used for question answering by formulating DL queries as below:
**Question 1**: Which cars are parked on roadside with trees behind and has some person sitting inside

[8] https://github.com/aspdr/VizITS/blob/main/VizITS.owl
[9] https://www.w3.org/TR/mediaont-10/

the Car and have width, height and area at least 50, 30 and 1000 ?
***DL Query***: **Car** and *parkedIn* some **Roadside** and *ahead* some **Tree** and *has* some **Person** and width min 50 and height min 30 and area min 1000
**Description**: Query returns the instances of class *Car* holding relation *parkedIn* with class *Roadside* and holds relation *ahead* with class *Tree* and holds the relation *has* with class *Person*. The instance should also have the width, height and area of atleast 50, 30, and 1000. The proposed ontology can represent spatial, temporal, occurrence-based actions, descriptive verbs, and prepositions, which are enough to represent all the possibilities occurring in the videos. Furthermore, video action representation, and verb descriptions, are highly domain-dependent, involving substantial domain information.

## 4 Conclusion and Future Scope

Within the scope of this thesis project, various strategies and methodologies are investigated with the goal of representing and identifying high-level video semantic information. This information can take the shape of events, activities, semantics, and general knowledge regarding the video data. A framework is described for event recognition and video interpretation with high-level semantics. This framework fills a well-known semantic gap that exists between the low-level properties of the video and the high-level (hidden) semantics of the video. Using semantic technologies, the reasoning is performed over mid-level and low-level features to generate high-level semantic information. To properly represent the video and conduct analysis on it, frame-level characteristics must first be taken from it. The selected frame-level information is mapped to data properties of the developed ontology. Using ad-hoc SWRL rules, such as object tracking rules, the relationships between objects in a video frame are determined. To evaluate how well the suggested method works in practice, a labeled dataset of suspicious actions is developed. The method is expanded to incorporate multi-modal analysis and data integration. A system that is capable of managing the data collected by mobile cameras and GPS sensors is also provided here. This framework makes use of an ontology-based method to integrate the data. Another integral part of this approach is ontology engineering, which requires substantial human resources, and is skilled in domain-specific concepts. Consequently, an NLP-guided data-driven approach constructs the *Class*, the *data-properties*, and the *object-properties* using the domain data, hence avoiding the requirement for expensive resources for ontology engineering. Unlike typical multimedia data, ontology-based data can express all semantic information in a structured and machine-readable fashion, saving a tremendous amount of space.

The depiction of video and the detection of events both bring up a wide variety of application cases and apply to several different fields. The majority of the actions that can be categorized as dangerous also include some everyday actions such as walking, running, standing, touching, sitting, opening the door, and other similar actions. These sub-activities do not require interaction between a number of items or the time domain; rather, they solely concentrate on a select number of objects. Identifying these sub-activities allows for the derivation of more complicated (and atypical) activities. Therefore, establishing and recognizing the visual relations present in the video data will be required in order to examine the process of identifying the more sophisticated activities that should be investigated.

## 5 Organization of the Thesis

The thesis contains one chapter for background and related work each, four contributory chapter of proposed work followed by conclusion and future scope. Furthermore, the brief description of each chapter is defined as following:
**Chapter-1 Introduction**: Presents the basic introduction and state-of-the-art architectures in Intelligent video surveillance systems with the overview of semantic technologies. Additionally, the motivation behind the work, research objectives, thesis organization, and major contributions have also been detailed.

**Chapter-2 Background and Related Work**: Provides background and related work about the current

state-of-the-art techniques for intelligent video surveillance system with emphasis on low-level features, mid-level features, and high-level video semantics.

**Chapter-3 Mid-level feature extraction for video representation and activity detection**: Presents the approach for mid-level feature extraction including object tracking, loitering, physical distancing and actual physical distance.

**Chapter-4 Semantic Technology for Video representation and analysis**: Describes methodology to represent video in machine-readable and machine-interpretable form using semantic technologies involving the low-level feature extraction and representation using developed ontology.

**Chapter-5 Ontology-based data integration and multi-modal analysis**: Describes ontology-based approach for data integration and multi-modal analysis by representing the data using an ontology. The reasoning is performed using SWRL rules and event identified using SPARQL.

**Chapter-6 Data-driven and NLP-guided ontology generation**: Presents a data-driven ontology generation approach for video scene representation and also demonstrate the methods to automate the tedious task of ontology generation using data-driven methods.

**Chapter-7 Conclusion**: Contains the conclusion and future work. General summary, limitations of the presented works, and the related future works have been highlighted. domain.

## 6 Publications

*Journal*

1. Patel, Ashish Singh, Ranjana Vyas, O. P. Vyas, Muneendra Ojha, and Vivek Tiwari, *"Motion-compensated online object tracking for activity detection and crowd behavior analysis"*, The Visual Computer, Springer, 1-21, IF:2.6, (2022). https://doi.org/10.1007/s00371-022-02469-3
2. Patel, Ashish Singh, Ranjana Vyas, O. P. Vyas, and Muneendra Ojha, *"A study on video semantics; overview, challenges, and applications"*, Multimedia Tools and Applications, Springer, 81, 6849–6897, IF:2.7, (2022). https://doi.org/10.1007/s11042-021-11722-1
3. Patel, Ashish Singh, Giovanni Merlino, Dario Bruneo, Antonio Puliafito, O. P. Vyas, and Muneendra Ojha, *"Video representation and suspicious event detection using semantic technologies"*, Semantic Web, IOS Press, 12(3), 467-491, IF:3.5, (2021) DOI: 10.3233/SW-200393
4. Patel, Ashish Singh, Giovanni Merlino, Antonio Puliafito, Ranjana Vyas, O. P. Vyas, Muneendra Ojha and Vivek Tiwari, *"An NLP-guided ontology development and refinement approach to represent and query visual information"*, Expert Systems with applications, Elsevier, IF: 8.6 (Minor Revision-Submitted)
5. Patel, Ashish Singh, Vivek Tiwari, Muneendra Ojha and O. P. Vyas, *"Ontology-based multi-modal media representation and retrieval using mobile vehicle for Intelligent Surveillance System"*, Soft Computing, Springer, IF: 3.7 (Under Review)
6. Patel, Ashish Singh, Vivek Tiwari, Muneendra Ojha and O. P. Vyas, *"Storing, analyzing, and retrieving semantic information: Public Place Video Surveillance"*, Journal of Visual Communication and Image Representation, Elsevier, IF: 2.8 (Under Review)

*Conference*

1. Patel, Ashish Singh, Muneendra Ojha, Monika Rani, Abhinav Khare, O. P. Vyas, and Ranjana Vyas, *"Ontology-based multi-agent smart bike sharing system (SBSS)"*, In 2018 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 417-422. IEEE, 2018.
2. Patel, Ashish Singh, O. P. Vyas, and Muneendra Ojha, *"Vehicle Tracking and Monitoring in Surveillance Video"*, In 2019 IEEE Conference on Information and Communication Technology, pp. 1-6. IEEE, 2019.
3. Ashish Singh Patel, and Muneendra Ojha, O P Vyas, *"Accident Detection Using Time-Distributed Model in Videos"*, In Proceedings of Fifth International Congress on Information and Communi-

cation Technology, pp. 214-223. Springer, Singapore, 2021.

4. Ashish Singh Patel, and Muneendra Ojha, O P Vyas,, *"Pronunciation Similarity Matching Using Deep Learning"*, In Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications, pp. 305-314. Springer, Singapore, 2021.
5. Patel, Ashish Singh, Vivek Tiwari, Muneendra Ojha and O. P. Vyas, *"VizOPS: A data-driven ontology to represent public place surveillance data"*, (accepted in IEEE IATMSI, 21-23 Dec 2022, IIIT Gwalior)
6. Patel, Ashish Singh, Vivek Tiwari, Muneendra Ojha and O. P. Vyas, *"Abandoned Luggage Detection: An Ontology-based approach for unusual activity recognition"*, (submitted in UPCON, IEEE, 02-04 Dec, 2022, IIIT Allahabad)

## References


[1] A. Ben Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Systems with Applications*, vol. 91, pp. 480–491, 2018. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2017.09.029

[2] "Cisco visual networking index: Forecast and methodology 2016-2021," 2017. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

[3] V. Vijayakumar and R. Nedunchezhian, "A study on video data mining," *International journal of multimedia information retrieval*, vol. 1, no. 3, pp. 153–172, 2012.

[4] H. Rai, M. H. Kolekar, N. Keshav, and J. Mukherjee, "Trajectory based unusual human movement identification for video surveillance system," in *Progress in Systems Engineering*. Springer, 2015, pp. 789–794.

[5] A. B. Mabrouk and E. Zagrouba, "Spatio-temporal feature using optical flow based distribution for violence detection," *Pattern Recognition Letters*, vol. 92, pp. 62–67, 2017.

[6] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[7] A. Dehghan, S. M. Assari, and M. Shah, "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4091–4099.

[8] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[9] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[10] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 803–818.

[11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[12] J. Yu, Y. Lee, K. C. Yow, M. Jeon, and W. Pedrycz, "Abnormal event detection and localization via adversarial event prediction," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[13] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, no. 3, pp. 86–91, 2006.

[14] L. F. Sikos, "Vidont : a core reference ontology for reasoning over video scenes scenes *," *Journal of Information and Telecommunication*, vol. 2, no. 2, pp. 1–13, 2018.

[15] J. Hunter, "Adding multimedia to the semantic web: Building an mpeg-7 ontology," 2001.

[16] M. Y. Kazi Tani, A. Lablack, A. Ghomari, and I. M. Bilasco, "Events detection using a video-surveillance ontology and a rule-based approach," in *European Conference on Computer Vision*. Springer, 2014, pp. 299–308.

[17] M. Y. K. Tani, A. Ghomari, A. Lablack, and I. M. Bilasco, "Ovis: ontology video surveillance indexing and retrieval system," *International Journal of Multimedia Information Retrieval*, vol. 6, no. 4, pp. 295–316, 2017.

[18] F. Sobhani and U. Straccia, "Towards a forensic event ontology to assist video surveillance-based vandalism detection," *CEUR Workshop Proceedings*, vol. 2396, pp. 30–47, 2019.

[19] U. Markowska-Kaczmar and H. Kwaśnicka, "Deep learning—a new era in bridging the semantic gap," in *Bridging the Semantic Gap in Image and Video Analysis*. Springer, 2018, pp. 123–159.

[20] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[21] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," 2016.

[22] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1218–1225.

[23] C. Dicle, O. I. Camps, and M. Sznaier, "The way they move: Tracking multiple targets with similar appearance," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2304–2311.

[24] L. F. Sikos, *Description logics in multimedia reasoning*. Springer, 2017.

[25] S. Smeureanu and R. T. Ionescu, "Real-time deep learning method for abandoned luggage detection in video," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1775–1779.

[26] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[27] J. Ferryman, "PETS 2006 Benchmark Data," 2006. [Online]. Available: http://www.cvg.reading.ac.uk/PETS2006/data.html